2011 11th IEEE International Conference on Computer and Information Technology

Predicting the Virtual Temperature of Web-Blog Articles as a Measurement Tool for Online Popularity

Su-Do Kim
Center of U-Port IT Research and Education
Pusan National University
Busan, Republic of Korea
kimsd@pusan.ac.kr

Sung-Hwan Kim and Hwan-Gue Cho*
Dept. of Computer Science and Engineering
Pusan National University
Busan, Republic of Korea
{sunghwan, hgcho}@pusan.ac.kr

Abstract— A **Blog provides commentary, news, or content on a particular subject. The important part of many blogs is interactive format. Sometimes, there is a heated debate on a topic; any article becomes a political or sociological issue. However, users not pay much attention to most articles. So, how we can predict the popularity of articles in advance and what is a standard for popularity? In this paper, we propose a methodology to predict the popularity of an article. First, we use an analogy between the virtual temperature and the popularity of the on-line articles. Thus, we define four different types of discrete temperature scale, such as explosive, hot, warm, and cold, according to the number of reviews in the saturated state of the article. We are concerned with how to predict the final temperature of the submitted articles in the internet Web-blog space. An experimental data set was collected from the articles submitted to "SEOPRISE", a well-known political discussion blog in Korea that more than 50,000 users visit per day. The hit count is used as a factor to predict the popularity, analogous to the number of viewers in the popularity of movies. We calculated the saturation point using the variation of hit count over the lifetime. We derived a sound regression model to predict the popularity temperature of the subject article in terms of the hit counts at the saturation point via the correlation coefficient of hourly hit count and hit count of the saturation point. We can predict the popularity temperature of Internet discussion articles using the hit count of the saturation point with more than 70% accuracy, exploiting only the first 30 minutes' information. Because of low predictive value of explosive, the results of prediction were worse than we think. In the hot, warm, and cold categories, we can predict more than 86% accuracy from 30 minutes and more than 90% accuracy from 70 minutes.**

Keywords; Prediction, Popularity, Discussion Blog, Hit count

1. INTRODUCTION

The user has an infinite desire for free space of communication. Users try to create a new network constantly. Users take advantage of the Web 2.0 services environment that can exchange and share the global scale information conveniently. Thus, new communication appears in blogs social networks, etc [1,2]. A blog provides commentary, news, or content on a particular subject. The interactive format is an important part of many blogs. The user is producer and also consumer in the blog. Sometimes, there is a heated debate about a topic; any article becomes a political or sociological issue and very popular. However, few articles receive much attention from users [3,4].

What factors stimulate user interest? The choice of factors depends on different standpoints from one person to another and from one content to another. The number of viewers is the most important factor to explain the popularity of movies and TV. Then, we can select the hit count that indicates the number of viewers of an article. This is an objective and quantitative factor.

How can we predict the popularity of the content in advance and what is the standard of popularity? Several researchers examined the role of social dynamics in explaining and predicting popularity [5-11]. Most research focuses on the social network based on the relationships between users, friends, followers or fans. Most blogs do not provide a specific tool of the relationship between users. The predictive model produces a large error in predicting the exact value of popularity [5]. Users tend to give more attention to articles with high popularity than those with less. That is, they find and select what article is popular, not the exact value [6].

We analyze the user characteristics of a discussion blog and study a methodology to predict article popularity based on the following questions: (1) what are dynamics in the hit count (2) can we predict the timing of the popularity (3) when should we make the prediction (4) how does the model predict (5) how should we present the popularity.

The remainder of this paper is structured as follows. Section 2 outlines related work. Section 3 analyzes the characteristics of blogs. Section 4 describes our prediction methodology and gives our experiment results. Finally, we conclude this paper in Section 5.

2. RELATED WORK

Lerman studied social information processing that plays an important role in document recommendation and filtering, and evaluating the quality of documents on Digg. She proposed a mathematical model that describes the dynamics of collective voting, analyzing the behavior of users [7,8].

978-0-7695-4388-8/11 $26.00 © 2011 IEEE
DOI 10.1109/CIT.2011.104

Jamali and Rangwala studied a co-participation network between users and the popularity of online content using comment information from Digg. They proposed features to predict the popularity by calculating the Digg-score using comments and relationship classification [9].

Szabo and Huberman proposed a linear correlation model by logarithmic transformation on the popularity of online content based on measurements of early and later user access in Digg and YouTube [6].

Lee, Moon and Salamatian proposed a methodology to infer likelihood that content will be popular. They used observable metrics, applying the Cox proportional hazard regression model for a set of given explanatory factors, such as thread lifetime, number of comments, and hit count [5].

Most prediction is based on the analysis of the user relationship in the social network. There is much uncertainty in expressing the exact value, because multiple factors affect the popularity over time. Users tend to give more attention to highly popular content. Typically, they want to know what contents is popular, not its exact value. We expressed the popularity of content in four virtual temperature scales.

3. CHARACTERISTICS OF BLOG ARTICLES

We used the dataset to analyze the characteristics of the discussion blog from Jan 1, 2010 to Dec. 31, 2010 in a discussion forum of SEOPRISE and AGORA. These are famous political discussion blogs in Korea [14,15].

Table I shows the characteristics of SEOPRISE and AGORA. SEOPRISE shows 110 articles and AGORA presents 20 articles per one page.

TABLE I. THE CHARACTERISTICS OF WEB-BLOG SEOPRISE AND AGORA IN 2010

2010 year	*SEOPRISE*	*AGORA*
Number of articles	83,763	694,170
Number of authors	24,400	33,343
Average number of articles per day	230	1,133
Average number of pages per day	2	57
Average hit count per article	1,026	102
Maximize hit count per article	309,595	234,760
Standard deviation of hit count per article	4,078	1,315
Average number of articles per author	3	43

The average hit count in articles is about 1000 in SEOPRISE and about 100 in AGORA. The hit count of most articles is very low, because the standard deviation is very high. Figures 1 and 2 show the hit count histograms. There are approximately 18000 (23%) articles above the average hit count in Figure 1 and 53000 (8%) in Figure 2. There are more articles (84%) with a hit count less than 50 greater in AGORA. A considerable number of articles receive a high hit count.

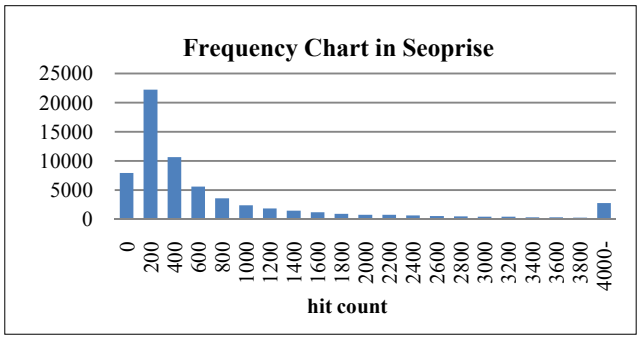

Figure 1. Frequency Hit Count Histogram for web-blog SEOPRISE: number of articles vs. hit count

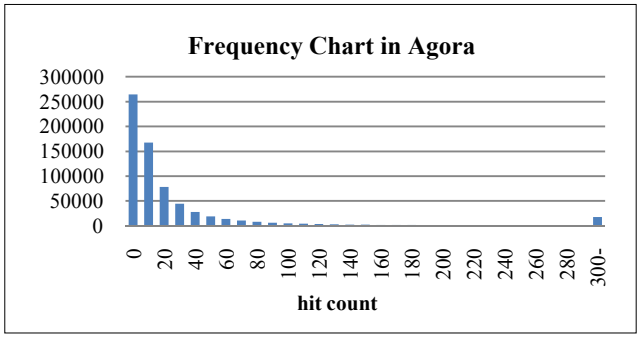

Figure 2. Frequency Hit Count Histogram for web-blog AGORA: number of articles vs. hit count

Figure 3 describes the hourly cycles in the rates of writing. The number of articles rapidly increase from 9AM and decrease from 6PM in both SEOPRISE and AGORA. The greatest growth in the number of articles is from 9AM to 12PM. The weekly cycle is a little different in SEOPRISE and AGORA. However, the number of articles is increased from Wednesday to Friday, and also decreased on weekends.

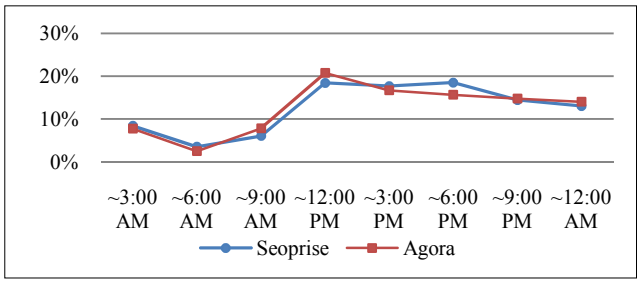

Figure 3. Hourly submission rates for the number of articles submitted in SEOPRISE and AGORA. The number of articles is averaged per 3 hours.

We tracked and analyzed the data of 310 articles from Mar. 15, 2011 to Mar. 18, 2011 in SEOPRISE. Figure 4 describes the variation on the average hit count for the first page. It shows a similar pattern to Figure 3. The user's activity pattern shows most users participate actively in discussion from 9AM to 6PM. Figure 5 describes the variation of the average hit count in pages.

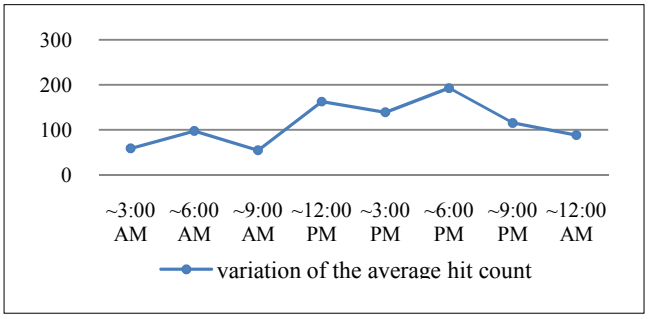

Figure 4. Hourly variation of the average hit count in 310 articles of SEOPRISE from Mar. 15, 2011 to Mar. 18, 2011.

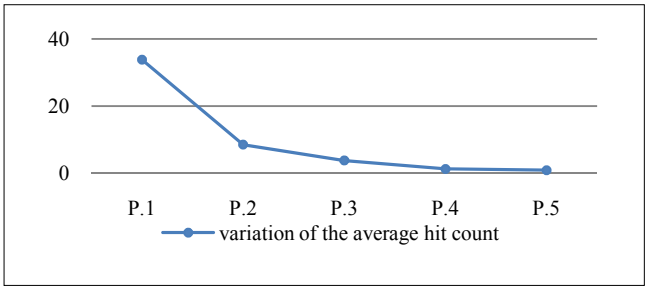

Figure 5. Page variation of the average hit count in 310 articles of SEOPRISE from Mar. 15, 2011 to Mar. 18, 2011.

The variations of the hit count significantly decrease when the page numbers increase. The variations of the hit count can be described by equation (1) for the page number of an article, as in (1). The residual standard error is 2.4592.

$$\Delta \mathrm{hit}(a_i, p) = 37.8875 \cdot p^{-2.3281} \quad (1)$$

4. Proposed Prediction Method

A. Data Set

We created a dataset from a discussion forum of SEOPRISE, a famous political discussion blog in Korea. Users submit on average 230 articles per a day. Approximately 2000 articles had a hit count above 5000 in 2010. The graphs show the hit count grows very soon after submission for most articles, but slows over time, and finally stops growing.

We made the dataset by tracking an article every 10 minutes after submitting the article in the forum. We subdivided the submission data into a training set and a test set to perform and validate the predictions. We took 816 articles submitted from March 15, 2011 to March 20, 2011, as the training set, and 1157 articles submitted from May 5, 2011 to may 10, 2011, as the test set.

Figure 6 shows tracking the hit count of 200 articles, where x is the time per 10 minutes.

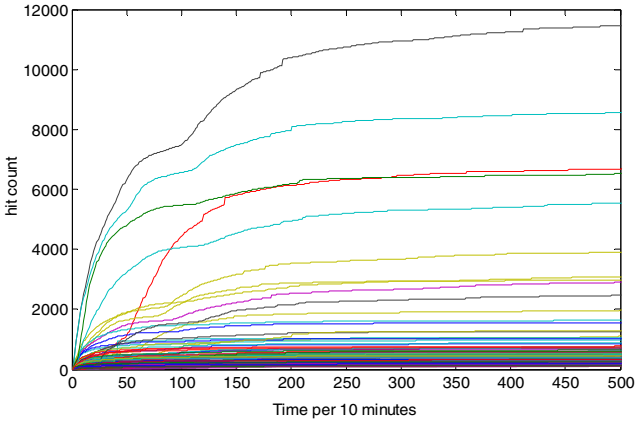

Figure 6. Hit count for 200 articles in SEOPRISE from March 15, 2011 to March. 20, 2011, we collected every 10 minutes.

B. Lifetime and Saturation point

Multiple factors affect the popularity of an article: hit count, comment, page, time, day of the week, author, title, and etc. The choice of factor determining the popularity of articles varies from one person to another. The number of viewers is the most important factor to explain the popularity for movies and TV. We choose hit count, an objective and quantitative value in blogs.

The number of viewers rapidly increases after the movie released, it slowly increases after a certain time, and finally the lifetime finishes. Articles show a similar pattern to movies. The hit count of an article grows very rapidly after submission in the early stages, then grows slowly, finally it stops growing. The popularity of a movie is mostly determined by the number of viewers from its release until it reaches the slow-growth stage [12]. We use the reference time at which we intend to predict the popularity to be the time when the growth rate falls. We term this time the "saturation point".

The dictionary definition of saturation point is the point at which the greatest possible amount of a substance has been absorbed. In economics, an individual demands a particular commodity due to the satisfaction or utility received from consuming it. Up to a point, the more units of a commodity the individual consumes per unit time, the greater the total utility received. Although total utility increases, the extra utility received from consuming each additional unit of the commodity usually decreases. At some level of consumption, the total utility received by the individual by consuming the commodity will reach a maximum; Marginal utility will be zero. This is the saturation point [13].

We define the lifetime and the saturation point as follows. We depict it in Figure 7.

Let us give two definitions:

a) Lifetime denotes the period from the submission to the time the variation of hit count exceeds a small value ε.

b) Saturation point denotes the time the variation of hit counts starts to decrease by more than the average variation.

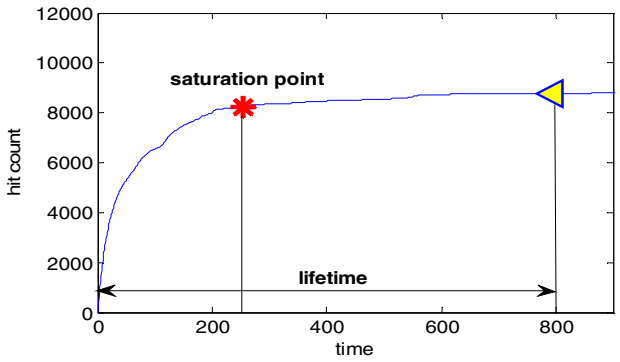

Figure 7. Lifetime and Saturation point

We define an article's lifetime as the interval until the variation of hit count no longer increases after submission. We denote this feature as $T_{life}(a_i)$ to an article a_i, as in (2). We set d during 24 hours until t_j and ε is a small value. We compared the mean of the variation for a day with ε, because the variation of hit counts is irregular. This is probably the users' access pattern, as shown in Figure 4.

$$T_{life}(a_i) = (t_o, t_j) \ , \ \sum_{k=j-d}^{j} \frac{\Delta hit(a_i, \Delta t_k)}{d} > \varepsilon \qquad (2)$$

The final hit count is the hit count at the end of the lifetime.

We term the saturation point as the starting time when the variation of the hit count begins to decrease. We denote this feature as $t_{sat}(a_i)$ for article a_i, as in (3). $\overline{\Delta hit}$ is the average variation of hit count except 0 during the lifetime.

$$t_{sat}(a_i) = t_j \ , \quad \sum_{k=j-d}^{j} \frac{\Delta hit(a_i, \Delta t_k)}{d} < \overline{\Delta hit} \qquad (3)$$

Figure 8 shows the lifetime and saturation point; these are now growing, if articles have no lifetime symbol .

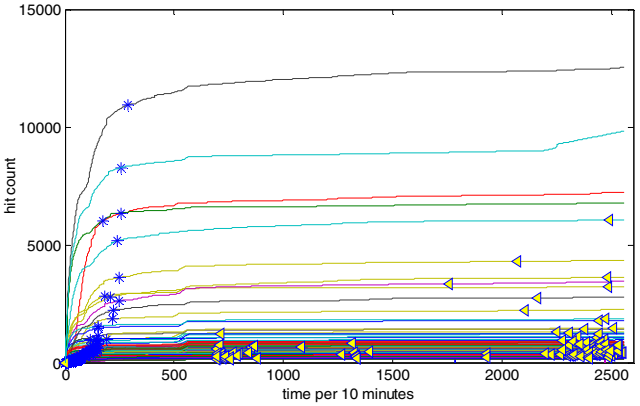

Figure 8. Symbols denote Lifetime(□) and Saturation point(*) of 200 articles collected from SEOPRISE.

We analyze the correlation of hourly hit count and the hit count of the saturation point to define the indicator time when performed the lifetime prediction. Table II shows the correlation coefficients of hourly hit count and hit count of the saturation point using the function of Pearson's correlation coefficient.

TABLE II. CORRELATION COEFFICIENTS OF HOURLY HIT COUNT AND SATURATION POINT HIT COUNT. THE CORRELATION IS STRONG AFTER 3TIME PERIODS (30 MINUTES).

R / t_j	*1*	*2*	*3*	*4*	*5*	*6*	*12*
$R(t_{sat}, t_j)$	0.49	0.69	0.70	0.70	0.70	0.71	0.71
$R(hit_{sat}, t_j)$	0.59	0.79	0.81	0.82	0.83	0.90	0.92

After 30 minutes, the hit count has the strongest correlation with the hit count of the saturation point above 0.8. We use the saturation point hit count to predict the popularity of an article, although it has the strongest correlation coefficient, 0.9966, between the hit count of the saturation point and the final hit count of lifetime. If an article is popular at the saturation point, it will be popular overall.

C. *Modeling popularity*

We transformed the hit count of 30 minutes and the hit count of the saturation point logarithmically and obtained a linear model with a strong linear correlation.

Figure 9 shows a linear model logarithmically transformed. We used the data except that much beyond the linear distribution. This can be described in equation (4). Figure 9 shows the predicted values represented by the green points.

$$hit_{sat}(a_i, t_{sat}) = a \cdot hit(a_i, t_j)^b \qquad (4)$$
$$\text{if } j = 3, \quad a = 0.4691 \ b = 1.5315$$

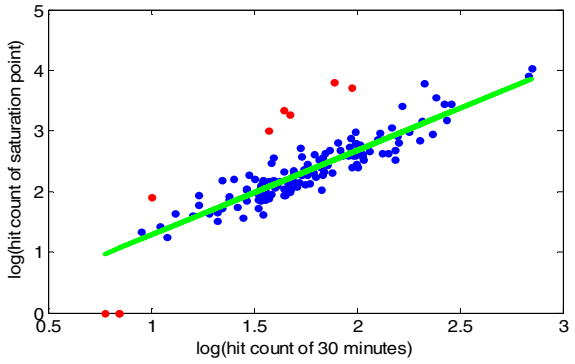

Figure 9. The Color depends on the distance to a linear model. The red point denotes articles far from the linear model, blue points are close to it.

The predictive value has many errors to express the exact value. The hit count does not increase equally due to multiple factors affecting the popularity of articles over their

lifetime [5]. Users tend to pay more attention to highly popular articles. Usually they want to know about popular articles.

Let us focus on measuring the popularity of posted articles. How do your separate the popularity of an article? In this paper, we use four types of virtual temperature based on hit count to express the popularity of an article.

$$\mathrm{Temp}(a_i) = \{ \text{explosive}, \text{hot}, \text{warm}, \text{cold} \}$$

- $\mathrm{Temp}(a_i)$=explosive, $1500 \le \mathrm{hit}_{sat}(a_i)$
- $\mathrm{Temp}(a_i)$=hot, $500 \le \mathrm{hit}_{sat}(a_i) < 1500$
- $\mathrm{Temp}(a_i)$=warm, $100 \le \mathrm{hit}_{sat}(a_i) < 500$
- $\mathrm{Temp}(a_i)$ =cold, $\mathrm{hit}_{sat}(a_i) < 100$

The standard used to separate hot, warm, cold is the hit count using the training set, after extracting the hit count of the top 20% and lower 10%. The explosive standard is the hit count of the top 5%.

D. Experiments

We chose 1157 articles submitted from May 4, 2011 to May 10, 2011, tracking the data every 10 minutes in SEOPRISE. We evaluated our proposed method using the predictive temperature model.

Table III shows the results of the predicted number of articles to the popularity temperature. 42, 69, 690, and 356 are the real number of articles to the respective popularity temperatures. The gap between the real number and predicted number lessens over time for the hot, warm, and cold values. However, the gap is very large for the explosive category, because the variation of hit count rapidly increases for a long time compared to the others and is hard to predict explosive category in advance.

TABLE III. EXPERIMENTAL RESULTS OF THE PREDICTED NUMBER OF ARTICLES TO POPULARITY TEMPARATURE OVER TIME. 20,33,215, AND 199 ARE THE REAL NUMBER OF ARTICLES.

T(10minutes)	***explosion***	***hot***	***warm***	***cold***
Actual number	42	69	690	356
t=1	18	54	934	151
t=2	6	49	861	241
t=3	8	61	783	305
t=4	7	63	786	301
t=5	8	66	775	308
t=6	10	62	771	314
t=7	10	62	757	327
t=8	11	64	744	338
t=9	11	63	748	335
t=10	12	61	742	342
t=11	14	59	726	358
t=12	13	59	726	359

Table IV shows the results of error rates for the real number of articles for each popularity temperature. The experiment shows the error rate is before 30 minutes. We calculated MAPE(Mean Absolute Percentage Error) is 43% at 10 minutes and 20 minutes. After 30 minutes, the MAPE is less than 30% and is 7% after two hours. Because of low predictive value of explosive, the results of prediction were worse than we think. The average correct rate is over 86% for the popularity temperature in the hot, warm, and cold categories from 30 minutes to 1 hour and over 90% after 70 minutes. Sometimes the error rate of a temperature increases and the error rate of the remaining temperatures decreased. The error rate is irregular, as the hit count does not increase at the same rate over time for diverse reasons. Table V show the examples of articles categorized into temperatures for SEOPRISE. No is the index number, date is the date of submission, and hit is the hit count.

TABLE IV. ABSOLUTED PE(PERCENTAGE ERROR) RATES OF THE PREDICTED NUMBER OF ARTICLES OF THE POPULARITY TEMPARATURES IN EXPERIMENTAL RESULTS OVER TIME.

T(10minutes)	***explosion***	***hot***	***warm***	***cold***
t=1	57%	22%	35%	58%
t=2	86%	29%	25%	32%
t=3	81%	12%	13%	14%
t=4	83%	9%	14%	15%
t=5	81%	4%	12%	13%
t=6	76%	10%	12%	12%
t=7	76%	10%	10%	8%
t=8	74%	7%	8%	5%
t=9	74%	9%	8%	6%
t=10	71%	12%	8%	4%
t=11	67%	14%	5%	1%
t=12	69%	14%	5%	1%

5. CONCLUSION

Some articles are very popular, while most articles are not. Users show an asymmetric attention to the discussion blog; a few articles receive the most attention, whereas most article s have a low hit count [6].

In this paper, we proposed a methodology to predict the popularity of an article using five approaches.

a) We used hit count as a factor to predict the popularity of an article.
b) We calculated the lifetime and the saturation point using an equation.
c) We determined the indicator time by a correlation coefficient of the hourly hit count and hit count of the saturation point.
d) We derived a model to predict the hit count of the saturation point.
e) We predicted the popularity temperature of an article (explosive, hot, warm, cold)

We collected 816 articles from Mar. 15, 2011 to Mar. 20, 2011, to derive a linear model. And we tested 1157 articles from May 4, 2011 to May 10, 2011, evaluating our proposed method using the temperature prediction model. We can predict over 86% of the popularity temperature of those in the hot, warm, and cold categories from 30 minutes to 60 minutes and over 90% after 70 minutes. The MAPE(Mean Absolute Percentage Error) is about 30% because the error

rate is very high in the explosive category. We can predict the popularity temperature of Internet discussion articles, using the hit count of saturation point with more than 70% accuracy by exploiting only the first 30 minutes information. In the hot, warm, and cold categories, we can predict more than 86% accuracy by exploiting only the first 30 minutes information.

Some articles have rapid growth, while others have slow growth in the early stage but may shows different patterns after. In this paper, we term such articles as "explosive". Predicting the popularity is very hard, and the error rate is high, because multiple factors affect the popularity of the article over its lifetime. We are interested in studying the effects of multiple factors on the popularity of articles and to analyze the characteristics of explosive articles.

TABLE V. EXAMPLES OF ARTICLES BASED ON POPULARITY TEMPARATURE IN SEOPRISE.

No.	Title	Author	Date	hit_{sat}	$hit_{t=3}$	Temp.
47560	3 communication companies	monsil	2011-05-10	32	8	cold
47553	Died together!	garlic	2011-05-10	588	126	hot
47528	Politicians' religion	donkey	2011-05-10	75	24	warm
47491	Leaving in silence	simsim	2011-05-10	42	6	cold
47371	The Democratic Party 34.5%	cowbell	2011-05-09	671	162	hot
47313	Don't you ruin yourself!	nicolle	2011-05-09	295	83	warm
46315	By-elections	bongha	2011-05-05	30	15	cold
46300	Lose an election in Kimhae	sad	2011-05-05	185	40	warm
46283	Internal divisions	clean	2011-05-05	39	15	cold
46136	Don't manipulate the hit count	anto	2011-05-04	7535	2688	explosive
46101	The way of the participation party, the way of Si-Min Rhyu	beautiful world	2011-05-04	1542	1109	explosive
46097	Why is opposite in the EU FTA?	hakill	2011-05-04	92	36	warm
46083	Si-Min Rhyu, what did you do wrong?	tong	2011-05-04	1705	1351	explosive
46065	Bad Si-Min Rhyu	old man	2011-05-04	1038	174	hot
46046	Here in korea	ramol	2011-05-04	128	61	warm
46023	Finally, Hak-Kyu Shon will come into Don-Young Chung	gone	2011-05-04	624	221	hot

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2010-371-B00008).

*Corresponding should be addressed to Hwan-Gue Cho (Email : hgcho@pusan.ac.kr)

REFERENCES

[1] Chih-Lu Lin and Hung-Yu Kao, "Blog Popularity Mining Using Social Interconnection Anaysis," IEEE Computer Society, vol 14, pp. 41-49, July 2010.

[2] N. Agarwal, H. Liu, L. Tang, and P.S. Yu, "Identifying the influential bloggers in a community," Proc. of the international conference on Web search and web data mining(WSDM), pp.207-218, 2008.

[3] Michaela Gőtz, Jure Leskovec, Mar.y McGlohon and Christos Faloutsos, "Modeling Blog Dynamics," Proc. of the Third International ICWSM Conference, pp. 26-33, 2009.

[4] Couronne Thomas, Stoica Alina and Beuscart Jean-Samuel, "Online social network popularity evolution: an additive mixture model," International Conference on Advances in Social Networks Analysis and Mining(ASONAM), pp. 346-350, 2010.

[5] Jong Gun Lee, Sue Moon and Kavé Salamatian, "An Approach to Model and Predict the Popularity of Online Conntents with Explanatory Factors," WI-IAT'10 Processings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, vol. 1, pp. 623-630, 2010.

[6] Gabor Szabo and Bernardo A. Huberman, "Predicting the Popularity of Online Content," Communication of the ACM, vol. 53, no. 8, pp. 80-88, August 2010.

[7] Kristina Lerman, "Social Information Processing in Social News Aggregation," IEEE Internet Computing:special issue on Social Search, vol. 11, no. 6, pp. 16-28, 2007.

[8] Kistina Lerman and Tad Hogg, "Using a Model of Social Dynamics to Predict Popularity of News," WWW '10 Proc. of the 19th international conference on World wide web, pp. 621-630, April 2010.

[9] Salman Jamali and Huzefa Rangwala, "Digging Digg: Comment Mining, Popularity Prediction, and Social Network Analysis," International Conference on Web Information Systems and Mining(WISM), pp. 32-38, 2009.

[10] Vicenc Gȯmez, Andreas Kaltenbrunner and Vicente Lȯpez, "Statistical Analysis of the Social Network and Discussion Threads in Slashot," International World Wide Web Conference Committee(WWW), pp. 645-654, April 2008.

[11] Andreas Kaltenbrunner, Vicenc Gȯmez and Vicente Lȯpez, "Description and prediction of slashdot activity," Proceddings of the 2007 Latin American Web Conference(LA_WEB), pp. 57-66, 2007.

[12] Sung an Ahn and Tae joon Kim, "Clustering by Life Cycle of Motion Picture," the Korean Journal of Advertising, vol 65, pp. 61-76, 2004.

[13] Dominick Salvatore, Schaum's outline of theory and problems of microeconomic theroy, 3rd ed., vol. 4. McGraw-Hill Professional, 1992, pp.63.

[14] SEOPRISE, http://www.seoprise.com/board/list.php

[15] AGORA,http://bbs1.agora.media.daum.net/gaia/do/debate/list?bbsId=D00